

Getting the Most Out of Multiple-choice Questions



David DiBattista, Ph.D.
Brock University
Professor of Psychology
3M National Teaching Fellow

Presented at
Wilfrid Laurier University
October, 2011

Contact information
David DiBattista, Ph.D.
Brock University
Department of Psychology
St. Catharines, ON L2S 3A1
E-mail: David.DiBattista@brocku.ca
Phone: 905-688-5550, ext. 3467

The Why and The How of Effective Testing

The primary goal of classroom testing

To measure the extent to which students have learned the facts, concepts, procedures, and skills that have been taught in the course.

An effective test

Students who have learned more will obtain higher test scores, and students who have learned less will obtain lower scores. To be effective, a test must consist of effective items.

What is an effective test item?

For a test item to be effective, students with higher test scores must be more likely to answer it correctly than students with lower scores.

		Percentage of students answering correctly	
		Bottom 25% of class	Top 25% of class
A good item			
A poor item			

Tips for Constructing Multiple-choice Items

- When writing the stem, use question format whenever possible and use sentence completion format only when absolutely necessary. In either case, use appropriate punctuation and capitalization.
- The stem should present the issue under consideration *clearly* and contain as much information as possible.
- Do not include irrelevant information in the stem unless it plays a role in the assessment procedure.
- Check carefully for spelling errors, giving special attention to distractors.
- If you use sentence-completion format, check carefully for grammatical consistency of stem and alternatives.
- Whenever possible, avoid negative wording in the stem, and be sure to emphasize it when it does occur.
- All distractors should be plausible.
- Four alternatives will usually be quite adequate, but the number used is best determined by the number of *plausible* distractors you can supply.
- To generate plausible distractors:
 1. Use students' most common errors on constructed-response tests.
 2. Use distractors that are similar to the correct answer in content, length, and complexity.
 3. Use words that sound important or have associations to the stem.
 4. Use distractors that are true, but do not correctly answer the question.
- Avoid patterns in the length and location of correct answers that could provide clues that are unrelated to content.
- Balance the answer key so that the correct response appears in each position about the same number of times.
- Do not use “none of the above.”
- Do not use “all of the above” unless there are only two distractors.
- Ignore any of the preceding suggestions when you have a good reason to do so.

Referring to “Tips for Constructing Multiple-Choice Items,” find at least one way in which each of the following items could be improved. Correct answers are italicized.

1. Sodium is the most abundant of the alkali metals, and the most common sodium compound is sodium chloride. A sodium chloride solution is said to be _____ when its concentration exceeds 0.9% (w/v).
 - A. hypertrophic
 - B. hyperplastic
 - C. hypertonic*
 - D. hypotonic

2. The Stanley Cup
 - A. Is made of cheese
 - B. Was made in England
 - C. Has been under the control of the National Hockey League since 1946, when there were only six teams in the league*
 - D. All of the above

3. Which of the following people was not a prime minister of Canada?
 - A. Pierre Trudeau
 - B. Sir John A. Macdonald
 - C. Terry Fox*
 - D. None of the above

Bloom's Taxonomy, Revised

COGNITIVE PROCESS DIMENSION

1. **REMEMBER:** Retrieving relevant knowledge from long-term memory
 - Recognizing; Recalling
2. **UNDERSTAND:** Determining the meaning of instructional messages, including oral, written, and graphic communications
 - Interpreting; Exemplifying; Classifying; Summarizing; Inferring; Comparing; Explaining
3. **APPLY:** Carrying out or using a procedure in a given situation
 - Executing; Implementing
4. **ANALYZE:** Breaking material into its constituent parts and detecting how the parts relate to one another and to an overall structure or purpose
 - Differentiating; Organizing; Attributing
5. **EVALUATE:** Making judgments based on criteria and standards
 - Checking; Critiquing
6. **CREATE:** Putting elements together to form a novel, coherent whole or make an original product
 - Generating; Planning; Producing

KNOWLEDGE DIMENSION

- A. **FACTUAL KNOWLEDGE:** The basic elements that students must know to be acquainted with a discipline or solve problems in it
 - Knowledge of terminology
 - Knowledge of specific details and elements
- B. **CONCEPTUAL KNOWLEDGE:** The interrelationships among the basic elements within a larger structure that enable them to function together
 - Knowledge of classifications and categories
 - Knowledge of principles and generalizations
 - Knowledge of theories, models, and structures
- C. **PROCEDURAL KNOWLEDGE:** How to do something; methods of inquiry, and criteria for using skills, algorithms, techniques, and methods
 - Knowledge of subject-specific skills and algorithm
 - Knowledge of subject-specific techniques and methods
 - Knowledge of criteria for determining when to use appropriate procedures
- D. **METACOGNITIVE KNOWLEDGE:** Knowledge of cognition in general as well as awareness and knowledge of one's own cognition
 - Strategic knowledge
 - Knowledge about cognitive tasks, including appropriate contextual and conditional knowledge
 - Self-knowledge

Bloom's Taxonomy, Revised

The Knowledge Dimension

The Cognitive Process Dimension

	A. Factual	B. Conceptual	C. Procedural	D. Metacognitive
1. Remember				
2. Understand				
3. Apply				
4. Analyze				
5. Evaluate				
6. Create				

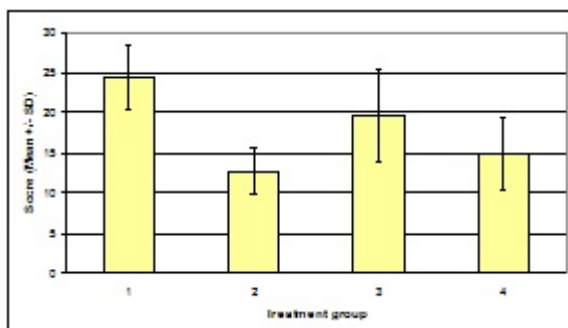
If assessment tasks are to tap higher-order cognitive processes, they must require that students cannot answer them correctly by relying on memory alone.

—Anderson and Krathwohl, 2001, page 71

Assessing Higher-level Cognitive Processes with Multiple-Choice Items

Understand: Interpreting

In the figure shown below, which treatment group has the most variability in its scores?



- A. Group 1
- B. Group 2
- C. Group 3*
- D. Group 4

Understand: Classifying

Right after a rat smells menthol, it is always given Drug X, which reliably induces substantial water intake. Eventually, the rat drinks water whenever it smells menthol, even when it is not injected with Drug X. In this situation, what is the role of Drug X?

- A. conditioned stimulus
- B. activational stimulus
- C. unconditioned stimulus*
- D. discriminative stimulus

Apply: Executing

Working with an ordinal data scale, Jeff obtained the following five scores: 0, 0, 2, 5, 8. What is the value of the median for this set of scores?

- A. 0
- B. 2*
- C. 3
- D. 5

Apply: Implementing

Working with an ordinal data scale, Jeff obtained the following five scores: 0, 0, 2, 5, 8. What is the value of the most appropriate measure of central tendency for this set of scores?

- A. 0
- B. 2*
- C. 3
- D. 5

Apply: Implementing

A researcher is planning an experiment in which each participant will receive both Treatment A and Treatment B. He is now thinking about the order in which the two treatments will be given to participants. Which of the following would be most advisable?

- A. Pick one order at random, either A-B or B-A, and give all participants the treatments in that order.
- B. For each participant, let the research assistant decide the order in which to give the treatments.
- C. Let participants decide for themselves the order in which they would prefer to be given the treatments.
- D. Give the treatments to half of the participants in the order A-B, and to the other half in the order B-A.*

Analyze: Differentiating

Keri received a grade of 70 on her history test. 200 people took the test, and scores ranged from a low of 30 to a high of 90. The class mean was 60, and the variance was 100. Which numbers must you use if you want to compute Keri's standard score?

- A. 30, 70, 90
- B. 30, 90, 200
- C. 60, 70, 100*
- D. 60, 100, 200

Analyze: Differentiating

A rat has been trained to press a bar to receive small pieces of food. The food dispenser has now been disconnected, but the rat is still hungry. Which of the following will have the greatest influence on how long the rat will continue to press the bar?

- A. the reinforcement schedule used during training*
- B. the total amount of food the rat ate during training
- C. the number of times the rat pressed the bar during training
- D. the amount of energy the rat expended to press the bar during training

Analyze: Organizing

Suppose that you are reviewing the history of the research that has been carried out on a particular topic over a very long period of time. Which of the following patterns would be most likely to characterize the methodological progress of the research?

- A. case studies first, then experimental studies, then correlational studies
- B. case studies first, then correlational studies, then experimental studies*
- C. experimental studies first, then case studies, then correlational studies
- D. experimental studies first, then correlational studies, then case studies

Analyze: Organizing

You place your finger in the middle of your forehead, and then move it straight back along the middle of your head and down the back of your head. Your finger started out near your frontal lobe and ended up near your cerebellum. Which parts of your brain did it come near along the way?

- A. first the parietal lobes, then the temporal lobes
- B. first the parietal lobes, then the occipital lobes*
- C. first the temporal lobes, then the parietal lobes
- D. first the temporal lobes, then the occipital lobes

Analyze: Attributing

During a session with a client, a therapist makes the following statement: “The problems that you are having now can be traced back to your relationship with your father when you were a little boy.” Which of these theorists has probably had the greatest influence on this therapist?

- A. Aaron Beck
- B. Carl Rogers
- C. B.F. Skinner
- D. Sigmund Freud*

Analyze: Attributing

Which of the following would a Rogerian therapist be most likely to say during a session with a client?

- A. You seem to be feeling a bit down today.*
- B. I want you to try to work things out with your sister.
- C. Your dream about going to the zoo—what do you think it might signify?
- D. There are some things I want you to work on before we meet again next week.

Analyze: Attributing

According to census data, people are having fewer children nowadays than they did 50 years ago. Your friend Anne tells you that she does not believe this because the young couple who live next door to her are both under 30 and already have four children. If Keith Stanovich were told about this, what might you reasonably expect him to say?

- A. The census data must be wrong.
- B. Anne’s comment illustrates valid probabilistic reasoning.
- C. Anne’s comment illustrates the use of person-who statistics.*
- D. The young couple provide an exception that actually serves to prove the rule.

Evaluate: Checking

You have carried out a 3×2 ANOVA for independent groups. There were 60 participants, with 10 participants randomly assigned to each cell. You have now analyzed the data and are double-checking your work. Which of the following would immediately let you know that you have made an error?

- A. You found the total degrees of freedom to be 60.*
- B. You found the mean square for the error term to be 6.25.
- C. You found the F -statistic for the interaction effect to be 2.34.
- D. You found degrees of freedom for the interaction effect to be 2.

Evaluate: Checking

Which of the following research findings would most strongly suggest that differences in Variable X are influenced by genetic factors?

- A. Sisters reared apart have more similar scores on Variable X than sisters reared together.
- B. Sisters reared together have more similar scores on Variable X than sisters reared apart.
- C. Same-sex fraternal twins reared together have more similar scores on Variable X than identical twins reared together.
- D. Identical twins reared together have more similar scores on Variable X than same-sex fraternal twins reared together.*

Evaluate: Checking

Which of the following research findings would most strongly support Schachter's theory of emotion?

- A. People injected with adrenaline find a slapstick movie funnier than those given a placebo.*
- B. Some people are much more expressive than others when they experience strong emotions.
- C. People who suffer damage to the amygdala frequently mistake expressions of anger for smiles.
- D. People who hold a pen in their teeth find cartoons funnier than people who hold a pen in their lips.

Evaluate: Critiquing

John is a healthy, well-adjusted young man who happens to become very nervous whenever he has to speak in public. To get over this problem, he starts taking some capsules that his doctor has told him will make him feel calmer when he speaks in public. John does not know it, but the capsules are actually empty. If John takes a capsule the next time he speaks in public, how do you think he will feel?

- A. more nervous than usual
- B. just as nervous as usual
- C. less nervous than usual*
- D. There is no way to predict how John will feel.

Evaluate: Critiquing

Dr. Brennan wants to compare the effectiveness of two training methods that are commonly used to teach people to juggle. He obtains a group of volunteers who have never juggled and randomly assigns each person to one of the two training methods. He sets alpha at 0.05 and determines that the power of his statistical test is 0.40. Which of the following is a valid criticism of Dr. Brennan's research study?

- A. Power is much too low; it should be at least 0.80.*
- B. Alpha is much too high; it should be no larger than 0.01.
- C. A placebo group should be included in the research design.
- D. Participants should be allowed to select the training method they will have.

Item Shells

Haladyna and Shindoll (1989) describe an item shell as a “hollow” MC item that has a syntactic structure, but no content. The test writer can insert important concepts into the item shell to construct challenging MC items. The use of item shells can make writing MC items easier and assist the writer in constructing challenging items. The following item shells are adapted from Haladyna (1997 and 2004; see *Further Reading on Multiple-Choice Testing* for further details).

Which best defines X?

Which is the meaning of X?

Which is synonymous with X?

Which is like X?

Which is characteristic (or uncharacteristic) of X?

Which is a defining characteristic of X?

Which is an example of X?

Which statement best exemplifies the principle of X?

Which is the cause of (or reason for) X?

Which is the relationship between X and Y?

A is to B as C is to which of the following?

Which is an example of the principle of X?

If X occurs, which is most likely to be the result?

Which is most commonly the cause of X?

Which distinguishes X from Y?

Which is most (or least) important (or significant, effective, etc.)?

Which is best (or worst, or highest/lowest, biggest/smallest, etc.)?

Which is most(or least) X?

Which is a difference (or similarity) between X and Y?

Which of the following principles applies to evaluating X?

Which is the most important factor contributing to X?

Which is a major shortcoming of X?

Problem presented. Which procedure (or strategy) would be used to solve this problem?

Problem presented. Which is a possible solution?

Problem presented. Why is X the most effective (or efficient) solution?

Context-dependent Item Set #1

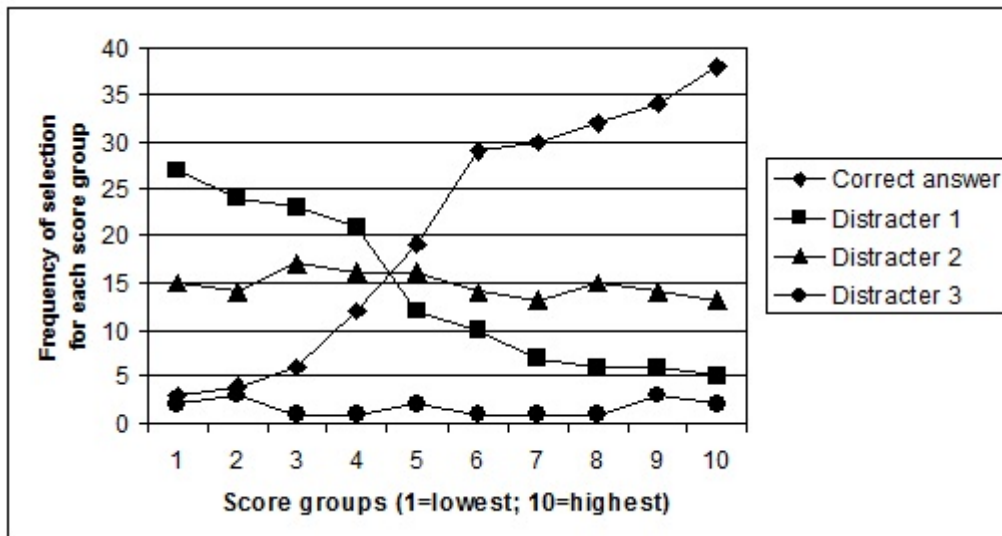
Instructions: Read the following research scenario and answer the questions that follow.

Madame Clousseau claims to be a psychic—that is, she claims to be able to predict future events with a level of accuracy better than chance. To examine her claim, Professor Jones brings her into his laboratory and tests her under carefully controlled conditions. He tosses a standard, fair coin 300 times and has Madame Clousseau predict what the outcome will be for each toss. He finds that she correctly predicts the outcome for 157 of the tosses. When he carries out the statistical test to analyze the results, Professor Jones lets alpha equal 0.05 and he uses a two-tailed test.

1. What statistical test should Professor Jones use to analyze the data?
 - A. the t -test for independent samples
 - *B. the z -test for binomial probability
 - C. the one-sample z -test
 - D. the one-sample t -test
2. What is the null hypothesis for the statistical test?
 - A. $\mu_0=150$
 - B. $\mu_0=157$
 - C. $\mu_1 - \mu_2=0$
 - *D. $\pi=0.50$
3. Suppose that the null hypothesis is actually **true**. What is the probability that Professor Jones will make a Type I error?
 - *A. 0.05
 - B. 0.10
 - C. 0.90
 - D. 0.95
4. Suppose now that the null hypothesis is actually **false**. If Professor Jones tossed the coin 50 times rather than 300 times, what effect would this have on the power of the statistical test?
 - A. The power would increase.
 - *B. The power would decrease.
 - C. The power would not be affected at all.
 - D. This question cannot be answered using the available information.
5. Suppose once again that the null hypothesis is actually **false**. If Professor Jones set alpha at 0.10 instead of at 0.05, what effect would this have on the power of the statistical test?
 - *A. The power would increase.
 - B. The power would decrease.
 - C. The power would not be affected at all.
 - D. This question cannot be answered using the available information.

Context-dependent Item Set #2

Instructions: The trace lines in the graph shown below were derived from a four-option multiple-choice item. Answer the questions that follow.



- Which of the following characteristics of the correct answer is particularly desirable?
 - It is selected more frequently than any of the distractors.
 - It is selected more frequently than all of the distractors put together.
 - *C. High-scoring students are more likely than low-scoring students to choose it.
 - D. Students with very little knowledge of the topic are likely to answer correctly by guessing.
- How would Distracter 1 be characterized?
 - plausible and a poor discriminator
 - *B. plausible and a good discriminator
 - C. implausible and a poor discriminator
 - D. implausible and a good discriminator
- How would Distracter 2 be characterized?
 - *A. plausible and a poor discriminator
 - B. plausible and a good discriminator
 - C. implausible and a poor discriminator
 - D. implausible and a good discriminator
- Of the following values, which is *closest* to the total number of students that selected Distracter 2?
 - A. 25
 - B. 50
 - C. 100
 - *D. 150

Context-dependent Item Set #3

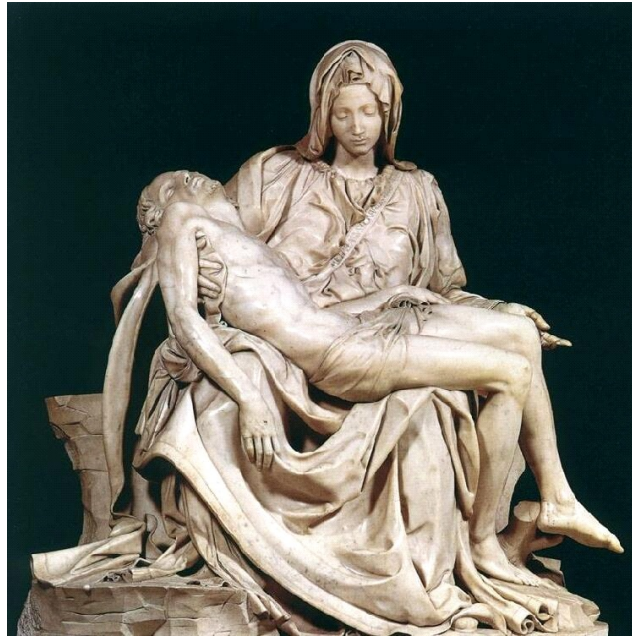
Instructions: Fill in all of the values missing from the following ANOVA summary table. Each correct answer earns one point.

Source of variation	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Between cells				
Variable A	600		300	
Variable B		1	100	
A×B				20.0
Within cells	420		10	
Total				

When you have completed the ANOVA table, answer the questions that follow. Each correct answer earns three points.

- What type of ANOVA was used to generate this table?
 - two-way repeated-measures ANOVA
 - *B. two-way independent-groups ANOVA
 - C. two-way between-within ANOVA
 - D. two-way mixed-design ANOVA
- What is the total number of scores in the data set that was used to generate this table?
 - *A. 48
 - B. 52
 - C. 53
 - D. 104
- How many levels of Variable A are there?
 - A. 1
 - B. 2
 - *C. 3
 - D. This question cannot be answered using the available information.
- What is the decision rule that relates to the *F*-statistic that has the **smallest** value in this table?
 - A. If $F \geq 3.22$, reject the null hypothesis.
 - B. If $F \geq 3.46$, reject the null hypothesis.
 - *C. If $F \geq 4.07$, reject the null hypothesis.
 - D. If $F \geq 5.07$, reject the null hypothesis.

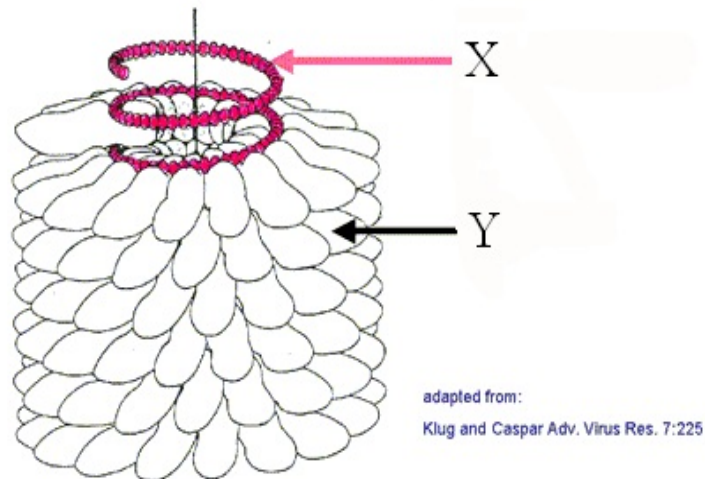
Context-dependent Item Set #4



Pieta

Michelangelo Buonarotti (1475-1564)

Context-dependent Item Set #5



Tobacco mosaic virus. X=nucleic acid; Y=protein.

The Cool Web

Children are dumb to say how hot the day is,
How hot the scent is of the summer rose,
How dreadful the black wastes of evening sky,
How dreadful the tall soldiers drumming by.

But we have speech, to chill the angry day,
And speech, to dull the rose's cruel scent.
We spell away the overhanging night,
We spell away the soldiers and the fright.

There's a cool web of language winds us in,
Retreat from too much joy or too much fear:
We grow sea-green at last and coldly die
In brininess and volubility.

But if we let our tongues lose self-possession,
Throwing off language and its watery clasp
Before our death, instead of when death comes,
Facing the wide glare of the children's day,
Facing the rose, the dark sky and the drums,
We shall go mad and die that way.

Robert Graves (1895-1985)

Instructions: Choose the best answer for each of the following items.

1. The poetic form of "The Cool Web" is best characterized as
 - A. free verse with a concluding rhymed couplet.
 - B. ballad stanzas with irregular rhymes.
 - C. blank verse with unusually irregular rhythm.
 - D. partially rhymed quatrains with a concluding sestet.
2. To emphasize the adult loss of childhood experience, the speaker of the poem
 - A. relies on frequent breaks in the middle of the line.
 - B. establishes a tone of caution, nostalgia and forgetfulness.
 - C. uses imagery of drowning.
 - D. alludes to the classical myth of endless return.

Continued on next page...

3. The “web of language” is *cool* because, according to the poem, language
 - A. is the means by which heated conflict may be resolved.
 - B. lessens the likelihood of achieving spiritual vision.
 - C. makes our register of the world less intense.
 - D. entangles us in misunderstandings.
4. When the speaker says that children “are dumb,” he means that they
 - A. experience life directly without the mediation of speech.
 - B. relate to the world with imagination rather than intellect.
 - C. lack the powers of rationality to comprehend the nuances of life.
 - D. become easily overpowered by the strength of their emotions.
5. According to the speaker, which of these is a time when adults might throw off the web of language?
 - A. At the moment of heightened passion.
 - B. At the moment of death.
 - C. At the moment of insight.
 - D. At the moment of belief.
6. Which of the following, according to the speaker, would be a consequence if adults were to “[throw] off language”?
 - A. Despair
 - B. Absurdity
 - C. Death
 - D. Madness
7. Since the poem decries the limitations of language, it is paradoxical that the speaker should rely on language so effectively. Arguably, this paradox is resolved by all of the following features—**except one**. Which one does **NOT** belong?
 - A. The speaker is speaking as an adult anyway, so the poem readily admits its intellectualization.
 - B. The speaker uses relatively simple and direct language, so he can hardly be accused of indulging in “volubility.”
 - C. The tone of the poem is unpretentious, and by using “we,” the speaker implicitly acknowledges his limitations.
 - D. The speaker makes a special claim for “the poet,” who is closer to the direct and emotional experiences of the child.

...etc.

Courtesy of Professor Brent MacLaine, Department of English, University of Prince Edward Island. Professor MacLaine uses MC items such as these in a team-learning context, with teams consisting of 4-6 students. Before coming to class, students read and study a piece of literature, such as “The Cool Web.” When they come to class, students first take a MC individually and submit it for grading. They then retake the same test, but this time working as a team. Team members must work out any disagreements they may have, and they must select one single response for each item. The team then submits a single answer sheet for grading. For each student, the higher of the two marks—individual test or team test—is the one that counts. Later, the entire class engages in an extended discussion of the poem, building on the insights gained in the team-learning setting.

Tips for Using Context-dependent Item Sets

In a context-dependent item set, a number of MC items follow the presentation of novel introductory material, such as a reading, scenario, data set, chart, or map. The MC items can be answered correctly only by referring to the introductory information. When using context-dependent item sets, keep the following in mind.

- Prepare stimulus material appropriate for the course in which it is to be used.
- Be certain that stimulus material is novel.
- Make stimulus material as brief and clear as possible.
- When creating MC items, follow the usual item-writing guidelines.
- Construct items that assess higher-order cognitive processes, not memory or fact-finding skill.
- Let the number of MC items be proportional to the length of the stimulus material.
- Context-dependent item sets can be used not only with MC items, but also with other response formats, including constructed-response formats.
- To reuse an item set, retain the form of an earlier item set and insert new content. Then make whatever minimal changes are necessary to harmonize the MC items with the new stimulus.

Adapted from Linn and Gronlund (1995)

Further Reading on Multiple-Choice Testing

These books focus specifically on multiple-choice testing.

- Haladyna, T.M. (2004). *Developing and Validating Multiple-Choice Test Items*, 3rd edition. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- McDonald, M.E. (2007). *The Nurse Educator's Guide to Assessing Learning Outcomes*, 2nd edition. Sudbury, MA: Jones and Bartlett.

These books provide background on assessment techniques in general and also have helpful information specifically about multiple-choice testing.

- Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of Educational Measurement*, 5th edition. Englewood Cliffs, New Jersey: Prentice-Hall.
- Linn, R.L. & Gronlund, N.E. (1995). *Measurement and assessment in teaching*, 7th edition. Upper Saddle River, New Jersey: Prentice-Hall.

This book focusses specifically on assessing higher-order thinking.

- Haladyna, T.M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Boston: Allyn and Bacon.

These journal articles examine the guidelines for writing multiple-choice items.

- Haladyna, T.M., & Downing, S.M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2, 37-50.
- Haladyna, T.M., & Downing, S.M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2, 51-78.
- Haladyna, T.M., Downing, S.M., and Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15, 309-344.

A novel response technique: The Immediate Feedback Assessment Technique

This journal article presents the Immediate Feedback Assessment Technique (IFAT), a novel multiple-choice response form that uses an answer-until-correct format and gives students immediate, corrective, item-by-item feedback while they take the test. Research has shown that the IFAT promotes learning and that students strongly prefer it to other response formats, such as the more widely used Scantron form. The following article provides further information and practical tips for instructors who might like to try the IFAT:

- DiBattista, D. (2005). The Immediate Feedback Assessment Technique: A learner-centered multiple-choice response form. *Canadian Journal of Higher Education*, 35, 111-131.

IFAT forms can be obtained from Epstein Educational Enterprises. If you wish to purchase forms or would like more information, go to <www.epsteineducation.com>.